Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

MS-E2115
Experimental and Statistical Methods in Biological Sciences

Lecture 8: Logistic regression

Joni Virta

Contents

1 Introductory example

2 Simple logistic regression
- Simple logistic regression model
- Parameter interpretation
- Diagnostics

3 Multiple logistic regression

4 References

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

Introductory example

Cautionary example

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- A dataset contains the records (sex, age, fare, survived or not) of 714 passengers onboard Titanic.
- We are interested in studying the relationship between survival (response) and sex, age and fare (explanatory variables).

	Survived	Sex	Age	Fare
1	0	male	22.00	7.25
2	1	female	38.00	71.28
3	1	female	26.00	7.92
4	1	female	35.00	53.10
5	0	male	35.00	8.05
6	0	male	54.00	51.86

Table: First 6 subjects of the Titanic dataset

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

Cautionary example

	0	1	Sum
female	8.96	27.59	36.55
male	50.42	13.03	63.45
Sum	59.38	40.62	100.00

Table: Cross-tabulation of Sex vs. Survived

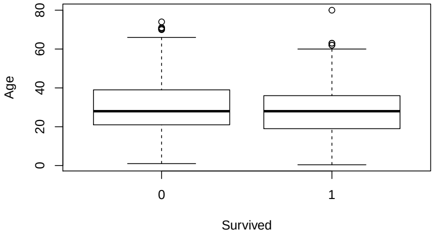


Figure: Boxplots of Age by Survived.

Cautionary example

Introductory example

Simple logistic regression

Simple logistic regression model

Parameter interpretation

Diagnostics

Multiple logistic regression

References

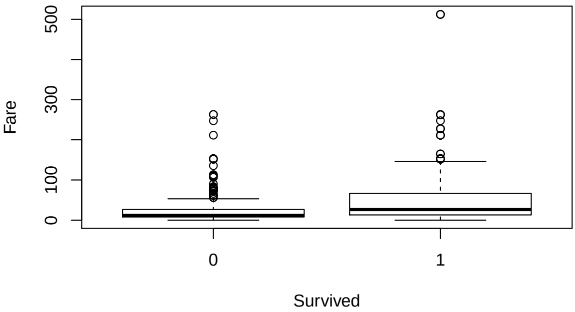


Figure: Boxplots of Fare by Survived.

Cautionary example

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- The response variable y_i (Survived) is binary and follows the Bernoulli distribution,

$$P(y_i = 1) = p_i = \text{prob. that passenger } i \text{ survives}$$
$$P(y_i = 0) = 1 - p_i = \text{prob. that passenger } i \text{ does not survive.}$$

- Recall that in linear regression, the expected value of the response, $E(y_i)$, is modelled with a linear function of the predictors,

$$E(y_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}.$$

- For Bernoulli distribution, $E(y_i) = p_i$ so we fit the line,

$$p_i = b_0 + b_1 \cdot (\text{sex})_i + b_2 \cdot (\text{age})_i + b_3 \cdot (\text{fare})_i.$$

Question:

What can go wrong with this approach?

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

Simple logistic regression

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

Simple logistic regression model

Link functions

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- The non-matching ranges of the two sides of the model equation are generally unified through the use of a link function, g, by assuming that:

$$g(E(y_i)) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip},$$

  for some g that transforms the range of the left-hand side to match that of the right-hand side.
- Popular ones to use when $E(y_i) \in [0, 1]$ are:
  1. The logit link: $g(p) = logit(p) = \log[p/(1-p)]$,
  2. The probit link: $g(p) = \phi^{-1}(p)$ (the quantile function of the standard normal distribution),
  3. The cloglog link: $g(p) = \log[-\log(1-p)]$.
- Link function is simply a way of changing the scale of the (expected value of the) response.

Assumptions

Introductory example

Simple logistic regression

Simple logistic regression model

Parameter interpretation

Diagnostics

Multiple logistic regression

References

The standard logistic regression is obtained by using the logit link.

Simple logistic regression, assumptions

- Consider n independent observation pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ of (x, y). Assume, that the values y_i are observed values of a binary random variable y and assume, for simplicity, that the values x_i are non-random.
- Assume that the logit-transformed expected values $p_i = E(y_i)$ depend linearly on the value x_i:

$$logit(p_i) = b_0 + b_1 x_i, \quad i \in \{1, ..., n\},$$

where the regression coefficients b_0 and b_1 are unknown constants.

Logistic curve

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- Thus we are modeling the probability to "succeed" as a function of the explanatory variable.
- When the value of the explanatory variable x_i is varied, the probability to succeed changes according to the relation,

$$p_i = logit^{-1}(b_0 + b_1 x_i).$$

- The resulting relationship is not linear but *logistic* (sigmoid) shape (analogy for the regression line in linear regression).

Example logistic curves

Introductory example

Simple logistic regression

Simple logistic regression model

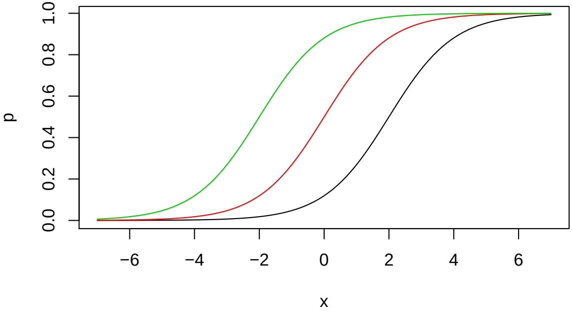
Parameter interpretation

Diagnostics

Multiple logistic regression

References

Logistic curves $logit^{-1}(b_0 + 1 \cdot x)$ for $b_0 = -2$, $b_0 = 0$ and $b_0 = 2$.



The intercept b_0 moves the logistic shape around x-axis.

Example logistic curves

Introductory example

Simple logistic regression

Simple logistic regression model

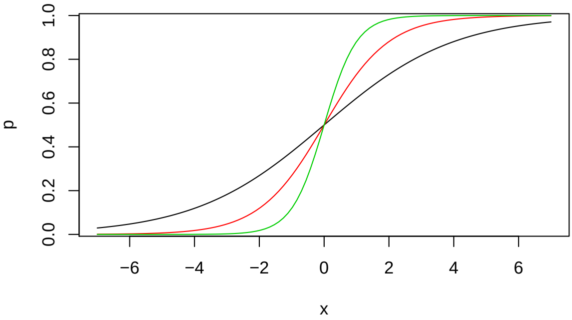
Parameter interpretation

Diagnostics

Multiple logistic regression

References

Logistic curves $logit^{-1}(0 + b_1 \cdot x)$ for $b_1 = 1/2$, $b_1 = 1$ and $b_1 = 2$.



The slope b_1 determines how steeply the probability of success grows with x_i.

Model fitting

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- As opposed to linear regression, the parameter estimates $\hat{b}_0$ and $\hat{b}_1$ do not have closed form expressions in logistic regression.
- The standard way of solving the logistic regression problem is through *iteratively weighted least squares* (IWLS).
- Most statistical software have logistic regression implemented in them.

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

Parameter interpretation

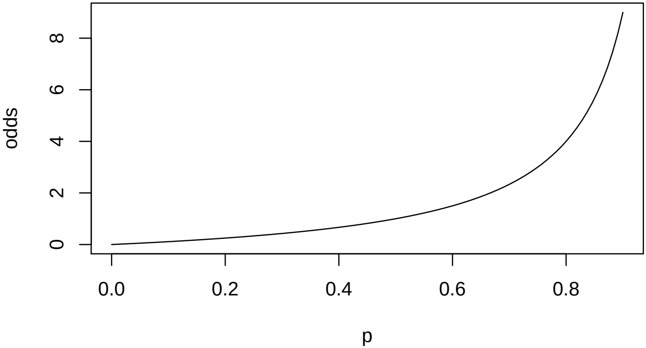
Odds

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- Odds is a way of writing probabilities used mostly in statistics and gambling.
- An event that has the probability p of happening is said to have odds (of happening) $odds(p) = p/(1-p)$.
- In layman usage this is usually written as "$1 : (1-p)/p$" or $p/(1-p) : 1$, depending on whether $p > 1/2$ or $p < 1/2$.
- Examples:
  - A probability $p = 1/2$ corresponds to even odds of $odds(p) = 1$ (written also as $1 : 1$)
  - A gambler wins the game with probability $p = 1/6$, or with odds $odds(p) = 1/5$ (written also as $1 : 5$).
  - The chance of rain is $p = 0.89$ or the odds of rain are $odds(p) = 8.09$ (written also as $8.09 : 1$)

Odds visually

The relationship between a probability p and the corresponding $odds(p) = p/(1-p)$.



Comparing odds

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- Odds ratios are used to compare the odds of two events with probabilities p_1 and p_2,

$$OR = \frac{odds(p_1)}{odds(p_2)}.$$

- Interpretation:
  - $OR < 1$: the second event is more probable than the first.
  - $OR = 1$: the two events are equally probable.
  - $OR > 1$: the first event is more probable than the second.

Odds ratio example

- A total of $p_1 = 42\%$ of subjects who received drug A had their condition improved and a total of $p_2 = 67\%$ of subjects who received drug B had their condition improved.
- The corresponding odds are $odds(p_1) = 0.724$ and $odds(p_2) = 2.030$
- The odds of the condition improving are 2.80 times higher for the subjects receiving drug B as compared to drug A.
- *Summary:* Odds ratio is simply a tool for comparing the probabilities of two events, e.g. the chance of survival under two different conditions.

Interpretation of the estimates

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- Odds ratios are closely connected to the interpretation of the regression coefficients in logistic regression.
- Recall, that in standard regression a change of one unit in a predictor causes a change in the expected value of the response equal to the corresponding regression coefficient.

$$E(y_i^*) - E(y_i) = (b_0 + b_1(x_i + 1)) - (b_0 + b_1 x_i) = b_1 .$$

- The same interpretation holds in the case of multiple predictors (assuming that the other predictors are held fixed and no interaction terms exist).

Interpretation of the estimates

Introductory example

Simple logistic regression

Simple logistic regression model

Parameter interpretation

Diagnostics

Multiple logistic regression

References

- In logistic regression, we have for the expected value $E(y_i) = p_i$

$$\log(odds(p_i)) = logit(p_i) = b_0 + b_1 x_i.$$

- Thus, for a change of one unit the predictor,

$$OR(p_i^*, p_i) = \frac{\exp(b_0 + b_1(x_i + 1))}{\exp(b_0 + b_1 x_i)} = \exp(b_1).$$

- The exponentiated coefficient describes the proportional change in odds corresponding to a single unit increase in the explanatory variable.
- If $b_1 > 0$, an increase in x will increase the odds (and probability) of "success" and vice versa.
- Note that this interpretation is only valid for the logistic link function.

Interpretation examples

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- In studying the connection between age (x_i) and a risk of a certain fictive disease (y_i), the following logistic regression model was fitted,

$$logit(p_i) = 1.23 + 0.03 \cdot x_i.$$

- *Interpretation:* Every year of age increases the odds of contracting the disease by a factor of $\exp(0.03) = 1.03$. E.g., the odds of contracting the disease are 35% higher for a 60 years old than for a 50 years old ($\exp(0.03)^{10} = \exp(0.30) = 1.35$).
- Note: if we use a qualitative predictor (e.g. 1 = female, 0 = male), then $\exp(b_1)$ describes the proportional change in odds for someone in class 1 vs. someone in class 0.

Inference

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- Again, the logistic regression model can always be fitted for any $0-1$ response, regardless whether the rest of the assumptions are fulfilled.
- However, the assumptions are required to make statistical inference on the parameters (compute confidence intervals, determine whether the parameters differ significantly from zero).
- The results are based on the central limit theorem and require large sample sizes.

Titanic, continued

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- We fit the model,

$$logit(P(Survived_i = 1)) = b_0 + b_1 Sex_i,$$

to the Titanic dataset using R, and obtain the following results:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1243	0.1439	7.81	0.0000
Sexmale	-2.4778	0.1850	-13.39	0.0000

- As in linear regression, the column `Pr(>|z|)` gives the p-value for the null hypothesis $H_0 : b = 0$ against the two-sided alternative and currently shows that both estimates differ significantly from zero.
- That is, the odds ratio of women vs. men surviving is $\exp(2.4778) \approx 11.9$ and this difference is statistically significant.

Titanic, continued

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- Simple approximate confidence intervals for the odds ratios are obtained by taking a confidence interval for the parameters and exponentiating them
- Approximate 95 % confidence interval for $-b_1$:

$$2.4778 \pm 1.96 \cdot 0.1850 = (2.1152, 2.8404).$$

- Approximate 95 % confidence interval for the odds ratio $\exp(-b_1)$ is:

$$(\exp(2.1152), \exp(2.8404)) = (8.2912, 17.1226).$$

- Similarly, approximate 95 % confidence interval for $\exp(b_1)$ is:

$$(\exp(-2.8404), \exp(-2.1152)) = (0.0584, 0.1206),$$

but the former is easier to interpret.

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

Diagnostics

Deviance

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- Logistic regression does not produce R-squared statistics or residuals in the same simple sense as the standard regression.
- Multiple different approaches for the two have been developed and the most common ones are based on the use of deviance, a generalization of sums of squares beyond ordinary regression.
- Using different forms of deviances (outputted by standard statistical software), the McFadden pseudo-R^2 is computed as,

$$\tilde{R}^2 = 1 - \frac{\text{ResidualDeviance}}{\text{NullDeviance}} \in [0, 1],$$

  with larger values indicating a better fit (0.20-0.40 can already be considered an excellent fit).
- For the titanic fit earlier, $\tilde{R}^2 = 0.222$.

Residuals and diagnostics

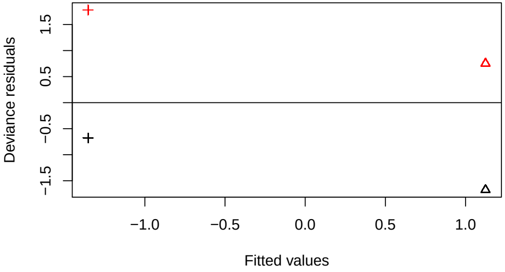
Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- The deviance can be decomposed into deviance residuals $\hat{\epsilon}$, $i = 1, \ldots, n$, an analogy for the standard residuals in ordinary linear regression.
- If the data is grouped (multiple observations have identical patterns of explanatory variables), the model assumptions can be checked (model diagnostics) by plotting the deviance residuals vs. the fitted values of the linear predictor $b_0 + b_1 x_i$.
- In the previous situation, if the model assumptions hold, the residuals,
  1. are approximately evenly distributed on both sides of zero,
  2. exhibit no unusual (non-linear) patterns in general.

Diagnostics for the Titanic dataset

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- Although the titanic dataset in the previous example is grouped, the small number of groups can make the previous conditions difficult to verify.

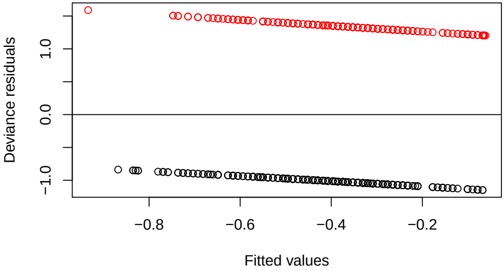


- The conditions seem to be verified...

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

Diagnostics for ungrouped data

- For ungrouped data, the previous diagnostic plot simply contains several “bands” of observations and has little value (though, outliers may be seen in the plot).



Overdispersion

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

- Unlike in the normal distribution, for Bernoulli distribution the variance $p_i(1-p_i)$ is a direct function of the expected value p_i.
- Thus for the assumption on Bernoulli-distributed responses to hold, no overdispersion (the variance of the observed responses is too large compared to their expected value) should occur.
- If no overdispersion has occurred the value,

$$\tilde{D} = \frac{\text{ResidualDeviance}}{n-q},$$

  where q is the number of parameters in the model, should be roughly around 1 (no more than $1+\sqrt{8/(n-q)}$)
- For the Titanic example, $\tilde{D} = 750.70/712 = 1.05$.

Causes of overdispersion

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

Overdispersion can be caused by

- missing explanantory variables,
- wrong link function,
- lack of non-linear effects,
- outliers,
- correlation between responses.

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

Multiple logistic regression

Introductory example

Simple logistic regression

Simple logistic regression model

Parameter interpretation

Diagnostics

Multiple logistic regression

References

Multiple logistic regression

- As with simple linear regression, also simple logistic regression can be extended to multiple logistic regression by simply adding more explanatory variables to the linear predictor.

Multiple logistic regression, assumptions

- Consider n independent observation pairs $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)$ of $(\mathbf{x}, y)$. Assume, that the values y_i are observed values of a binary random variable y and assume, for simplicity, that the values $\mathbf{x}_i$ are non-random.
- Assume that the logit-transformed expected values $p_i = E(y_i)$ depend linearly on the vector $\mathbf{x}_i$:

$$logit(p_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}, \quad i \in \{1, ..., n\},$$

where the regression coefficients $b_0, b_1, ..., b_p$ are unknown constants.

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

Multiple logistic regression

- Everything that was said about simple logistic regression (solutions, inference...) can be extended to multiple logistic regression but goes beyond the scope of this course.
- We simply go through an introductory example, the titanic dataset with all three predictors included.

	Survived	Sex	Age	Fare
1	0	male	22.00	7.25
2	1	female	38.00	71.28
3	1	female	26.00	7.92
4	1	female	35.00	53.10
5	0	male	35.00	8.05
6	0	male	54.00	51.86

Table: First 6 subjects of the Titanic dataset

Titanic, continued

- We fit the model,

$$logit(P(Survived_i = 1)) = b_0 + b_1 Sex_i + b_2 Age_i + b_3 Fare_i,$$

to the Titanic dataset using R, and obtain the following results:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9348	0.2391	3.91	0.0001
Sexmale	-2.3476	0.1900	-12.36	0.0000
Age	-0.0106	0.0065	-1.63	0.1038
Fare	0.0128	0.0027	4.74	0.0000

- The pseudo$-R^2$ and the overdispersion statistic equal,

$$\tilde{R}^2 = 0.258 \quad \text{and} \quad \tilde{D} = 1.009.$$

- Computing the VIFs does not reveal significant multicollinearity.

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

Final note

- Both logistic and ordinary linear regression are special cases of the so-called *generalized linear models* (GLM).
- GLM:s are characterized by the distribution of the response variable y:
  - Normal $\to$ linear regression,
  - Bernoulli $\to$ logistic regression,
  - Poisson $\to$ log-linear models,
  - Negative binomial $\to$ negative binomial regression.
- Similar results as seen on lectures 7 and 8 are available for other members of the GLM-family as well.

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

References

References

Introductory example
Simple logistic regression
Simple logistic regression model
Parameter interpretation
Diagnostics
Multiple logistic regression
References

F. E. Harrell, Jr.: Regression Modeling Strategies, Springer 2015.